Bioinspired Visuovestibular Artificial Perception System for Independent Motion Segmentation

Jorge Lobo, João Filipe Ferreira and Jorge Dias
Institute of Systems and Robotics
University of Coimbra — Polo II
3030-290 Coimbra, Portugal
{jlobo,jfilipe,jorge}@isr.uc.pt

Abstract

In vision based systems used in mobile robotics and virtual reality systems the perception of self-motion and the structure of the environment is essential. Inertial and earth field magnetic pose sensors can provide valuable data about camera ego-motion, as well as absolute references for structure feature orientations. In this article we present several techniques running on a biologically inspired artificial system which attempts to recreate the "hardware" of biological visuovestibular systems resorting to computer vision and inertial-magnetic devices. More specifically, we explore the fusion of optical flow and stereo techniques with data from the inertial and magnetic sensors, enabling the depth flow segmentation of a moving observer. A depth map registration and motion segmentation method is proposed, and experimental results of stereo depth flow segmentation obtained from a moving robotic/artificial observer are presented.

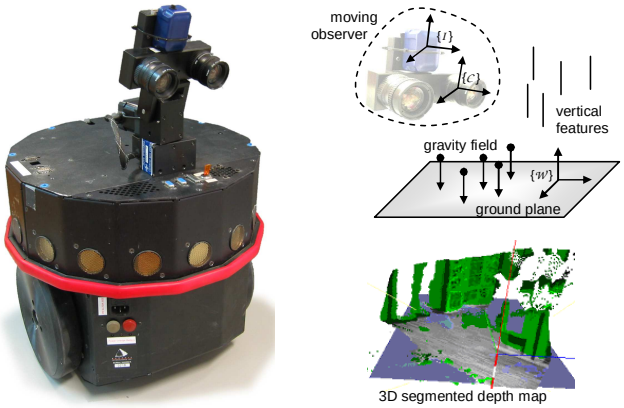

Figure 1. Stereo vision system with an inertial measurement unit used on robotic system, frames of reference and its 3D segmented depth map output.

1. Introduction

In biological vision systems, inertial cues provided by the vestibular system play an important role, and are fused with vision in the early processing stages of image processing (e.g, the gravity vertical cue). Artificial perception systems for robotic applications have since recently been taking advantage from low-cost inertial sensors for complementing vision systems, using both static and dynamic cues.

Inertial sensors attached to a camera can provide valuable data about camera pose and movement. Micromachining enables the development of low-cost single-chip inertial sensors that can be easily incorporated alongside the camera's imaging sensor, thus providing an artificial vestibular system. Figure 1 shows a stereo-camera pair with an inertial measurement unit (IMU) mounted on a mobile robotic platform. The 3D-structured world is observed by the visual sensor, and its pose and motion are directly measured by the inertial sensors. These motion parameters can also be inferred from the image flow and known scene features. Combining the two sensing modalities simplifies the 3D reconstruction of the observed world. Inertial sensors also provide important cues about the observed scene structure, such as vertical and horizontal references. Inertial navigation systems obtain velocity and position by integration, and do not depend on any external references, except gravity.

The inertial-sensed gravity vector provides a unique reference for image-sensed spatial directions. More specifically, previous work has shown that the use of visual sensors together with IMUs can be used to estimate camera focal distance [9] or to perform cross-calibration [3]. Knowing the vertical-reference and stereo-camera parameters, the ground plane can be fully determined. The collineation

between image ground-plane points can be used to speed up ground-plane segmentation and 3D reconstruction. Using the inertial reference, vertical features starting from the ground plane can also be segmented and matched across the stereo pair, so that their 3D position is determined. The inertial vertical reference can also be used after applying standard stereo-vision techniques; taking the ground plane as a reference, the fusion of multiple maps reduces to a 2D translation and rotation problem, and dynamic inertial cues may be used as a first approximation for this transformation, providing a fast depth-map registration method (Figure 1) [8]. In addition, inertial data can be integrated into optical flow techniques, through compensating camera egomotion, improving interest-point selection, matching the interest points, and performing subsequent image-motion detection and tracking for depth-flow computation. The image focus of expansion and centre of rotation are determined by camera motion and can both be easily found using inertial data alone, provided that the system has been calibrated. This information can be useful during vision-based navigation tasks.

Three-dimensional scene flow estimation was studied by Vedula *et al.* [15][14]. Several scenarios are presented, and the tradeoffs between structure knowledge, correspondence matching, number of cameras and computed optical flow explored. Dense scene flow estimation using only two cameras was proposed by Li and Sclaroff by fusing stereo and optical flow estimation in a single coherent framework [7]. Ye Zhang and Kambhamettu computed dense 3D scene flow and structure from multiview image sequences with non-rigid motion in the scene [17]. Stereoscopic MPEG based video compression methods also deal with motion flow segmentation, such as the joint motion and disparity fields estimation method proposed by Yang *et al.* [16]. A statistical approach to background modelling was used for segmentation of video-rate stereo sequences by Eveland *et al.*[5].

However, when dealing with a free moving stereo camera observer, the methods described above are not directly applicable. Visual and inertial sensing are two sensory modalities that can be explored to give robust solutions on image segmentation and recovery of 3D structure from images [9]; inertial sensors provide valuable data to deal with the camera motion [10]. Consequently, artificial systems dealing with motion perception in more complex situations would clearly gain by introducing *bionsinspired* visuovestibular sensing using computer vision and inertial-magnetic devices. In this article, we will present approaches for ego-motion and independent motion perception and segmentation based on these biologically inspired visuovestibular artficial systems.

Correlation-based stereo depth maps can be generated from a moving vision system, and rotated to a common levelled reference provided by the rotation update from inertial

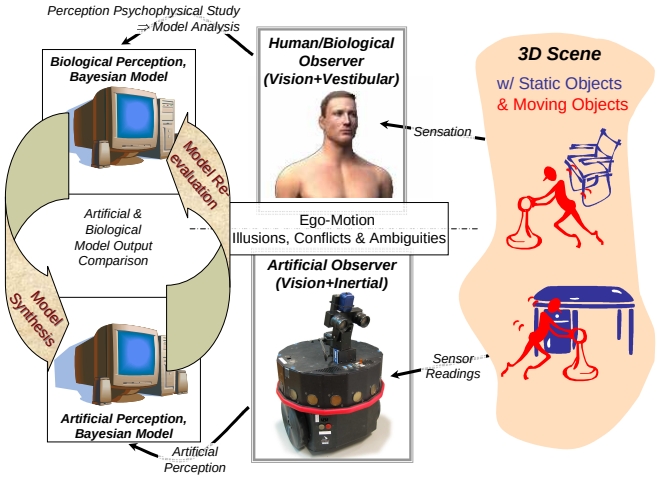

Figure 2. Biomimetic artificial perception research proposal schematic (human observer image courtesy of 3DScience.com).

sensed gravity and magnetic sensed bearing. Voxel quantization can then be performed on the resulting maps. But there remains a 3D translation in the successive depth maps due to the motion, for which the inertial sensors only provide a rough estimate. By tracking some image targets over successive frames, the system translation between frames can be estimated by subtracting their 3D position. The translation can also be estimated from the 3D data alone. For scenes where a base horizontal plane is always visible (eg: the floor or desktop), a histogram in height can be used to have a common reference along the vertical axis. This can also be performed for the horizontal axis if the orientation of visible planes is known or detected by a 2D fit to the data. The two identified planes provide the translation to merge successive depth maps.

Fully registered depth maps can therefore be obtained from the moving system — our solution for correlation-based stereo depth map registration is presented on section 2.

The depth flow that remains in the resultant map is due to the system covering new scenes, or to moving objects within the overlap volume of successive observations. Mismatches between the depth from stereo and depth from optical flow indicate possible independent motion. This can be used to better segment moving objects in the overlap volume and avoid artifacts from slow moving objects. On section 3 we describe our approaches for independent motion segmentation using these registered maps.

Subsequently, we present results on section 4, draw some conclusions on section 5 and finally, on section 6, we discuss the outcome of our studies and propose future work

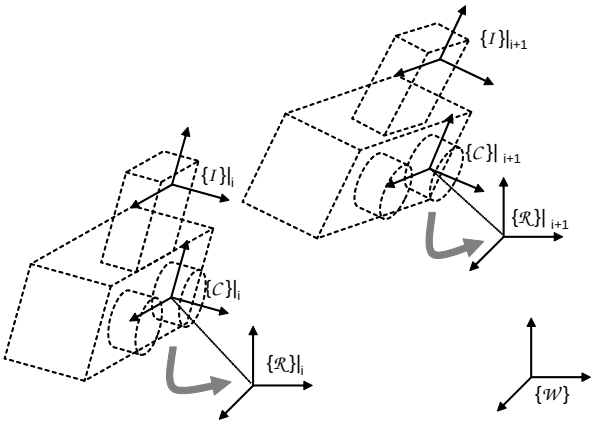

Figure 3. Moving observer and world fixed frames of reference.

which will pave way to the next step in this line of research: from *bioinspired* to *biomimetic*.

As can be seen on Figure 2, the outcome of this kind of research may become not only a significant advance in artificial sensing, such as in the solution of ego-motion or independent motion segmentation which we propose to tackle directly, but also a challenge in allowing further insight on aiding human beings to surpass their own perceptive limitations, helping disambiguation and coping with illusions or conflicts arising in extreme conditions where humans are prone to failure, namely: in extreme environments (e.g., in space exploration), where humans are displaced from normal conditions and factors such as 1G gravity; in perceptive pathologies (i.e. perception-impaired patients).

2. Registering Stereo Depth Maps

Biological visuovestibular systems take into account ego-motion, and deal well with independent motion segmentation, while they successfully integrate 3D information of the surrounding scene as the biological observer moves along its trajectory. Taking this into account, our approaches to ego and independent motion segmentation were devised so as to take advantage of the output of a preceding stereo depth map registration procedure.

A moving stereo observer of a background static scene with some moving objects can compute at each instant a correlation-based dense depth map. The maps will change in time due to both the moving objects and the observer ego-motion. A first step to process the incoming data is to register the maps to a common fixed frame of reference $\{\mathcal{W}\}$, as shown on Figure 3.

The stereo cameras provide intensity images $I_l(u,v)|_i$ and $I_r(u,v)|_i$, where u and v are pixel coordinates, and i the frame time index. Having the stereo rig calibrated, depth maps for each frame can be computed. A set of 3D points ${}^{\mathcal{C}}\mathbb{P}|_i$ is therefore obtained at each frame, given in the camera frame of reference $\{\mathcal{C}\}|_i$. Each 3D point has a corresponding intensity gray level c given by the pixel in the reference camera, i.e $c = I_l(u,v)|_i$. Each point in the set retains both 3D position and gray level

$$P(x,y,z,c) \in {}^{\mathcal{C}}\mathbb{P}|_i \ . \tag{1}$$

2.1. Rotate to Local Vertical and Magnetic North

The inertial and magnetic sensors, rigidly fixed to the stereo camera rig, provide a stable camera rotation update ${}^{\mathcal{R}}\boldsymbol{R}_{\mathcal{C}}$ relative to the local gravity vertical and magnetic north camera frame of reference $\{\mathcal{R}\}|_i$.

Calibration of the rigid body rotation between $\{\mathcal{I}\}|_i$ and $\{\mathcal{C}\}|_i$ can be performed by having both sensors observing gravity, as vertical vanishing points and sensed acceleration, as described in [11].

The rotated camera frame of reference $\{\mathcal{R}\}|_i$ is time-dependent only due to the camera system translation, since rotation has been compensated for.

2.2. Translation from Image Tracked Target

The translation component can be obtained using a single fixed target tracked in the scene. The image feature must have the corresponding 3D point P_t in each depth map, so that translation can be estimated from

$$\Delta\vec{t} = \boldsymbol{P}_t|_{i+1} - \boldsymbol{P}_t|_i \tag{2}$$

with $\boldsymbol{P}_t|_{i+1} \in {}^{\mathcal{R}}\mathbb{P}|_{i+1}$ and $\boldsymbol{P}_t|_i \in {}^{\mathcal{R}}\mathbb{P}|_i$.

The fixed target can be an artificial one, or set of sparse tracked natural 3D features can be used to improve robustness, but assumptions have to be made in order to reject outliers that occur from tracking features of the moving objects.

2.3. Voxel Quantization

The above equations are provided for discrete sets of points. In order to deal with noise and allow 3D volume processing, a 3D array is built representing 3D space as voxels. For each stereo frame, the corresponding cubic array of voxels $Vox|_i$ can be built. For the occupied voxels the corresponding gray level can be stored in the array. When two or more points contribute to the same voxel, the average gray level is used.

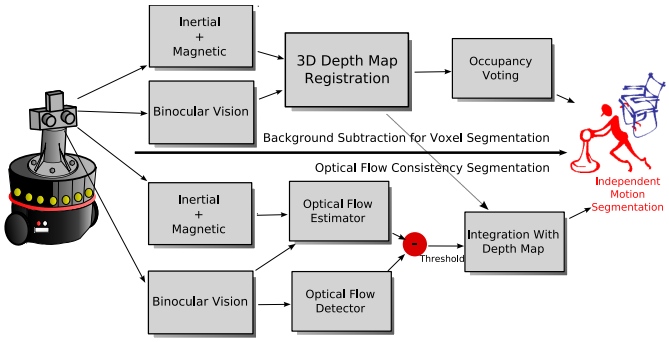

Figure 4. Summary of voxel background subtraction and optical flow consistency analysis methods for independent motion segmentation.

For each $P(x, y, z, c) \in {}^{\mathcal{C}}\mathbb{P}|_i$, $Vox(x, y, z)|_i = c$ if previously empty, or $Vox(x, y, z)|_i = \bar{c}$, where $\bar{c}$ is the average gray level of the contributing points.

For a sequence of stereo frames, two cumulative voxel arrays Vox_c and Vox_v can be built for both gray level and occupancy statistics over the frames, with

$$Vox_c(x, y, z) = \bar{c}_v, \qquad Vox_v(x, y, z) = v \tag{3}$$

where v is the number of frames that voted voxel (x, y, z) as occupied, and $\bar{c}_v$ the average gray level from the voting frames.

3. Independent Motion Segmentation in Fully Registered Maps

Having the dense depth maps in a common frame of reference we can proceed to segment the moving objects seen by the moving stereo observer. Biological vision systems are very successful in movement segmentation since they efficiently resort to flow analysis and accumulated prior knowledge of the 3D structure of the scene. Artificial perception systems may also build 3D structure maps and use optical flow to provide cues for ego and independent motion segmentation (see Figure 4). The maps will change in time due to moving objects, and eventually grow as the artificial observer covers new scene areas.

3.1. Background Subtraction for Voxel Segmentation

Occupancy statistics can be used to segment the set of voxels that correspond to the static scene observed by the moving system, and segment the moving objects.

Applying a threshold v_{back} on the accumulated vote count, a binary array of background voxels Vox_b can be built as

$$Vox_b(x, y, z) = 1 \text{ when } Vox_v(x, y, z) > v_{back} \ . \tag{4}$$

To improve noise filtering and robustness, a thinning and growing transformation is applied, removing isolated voxels and filling in gaps. The thinning filter takes out voxels without a minimum number of neighbours, by performing a convolution with a cubic unit kernel and thresholding the result back to a binary array. The growing simply performs a convolution with a cubic unit kernel, and rebuilds the binary array with all the non-zero voxels.

For a single frame i, the set of voxels from moving objects will be given by

$$Vox_m|_i = Vox|_i \bigcap \overline{Vox_b} \ . \tag{5}$$

To deal with noise, thinning and growth smoothing can also be applied to $Vox_m|_i$, but smearing of the intensity gray level might not help subsequent 3D intensity based methods.

The underlying assumption is that the moving observer repeatedly covers the same scene so that background voxels are seen more times than moving objects. Experimental results show that moving objects are successfully segmented and that thinning and growth smoothing filter out noise from the correlation based stereo depth maps.

3.2. Optical Flow Consistency Segmentation

Optical flow is the apparent motion of brightness patterns in the image. Generally, optical flow corresponds to the projected motion field, but not always. Shading, changing lighting and some texture patterns might induce an optical field different from the motion field. However since what can be observed is the optical field, the assumption is made that optical flow field provides a good estimate for the true projected motion field.

Optical flow computation can be made in a *dense* way, by estimating motion vectors for every image pixel, or *feature based*, estimating motion parameters only for matched features.

Representing the 2D velocity of an image pixel $\boldsymbol{u} = (u, v)^{\mathsf{T}}$ as $\frac{d\boldsymbol{u}}{dt}$, the brightness constancy constraint says that the projection of a world point has a constant intensity over a short interval of time, i.e., assuming that the pixel intensity or brightness is constant during dt, we have

$$I(u + \frac{du}{dt}dt, v + \frac{dv}{dt}dt)|_{t+dt} = I(u, v)|_t \tag{6}$$

Figure 5. Experimental setup of 3D scene with static background and swinging pendulum.

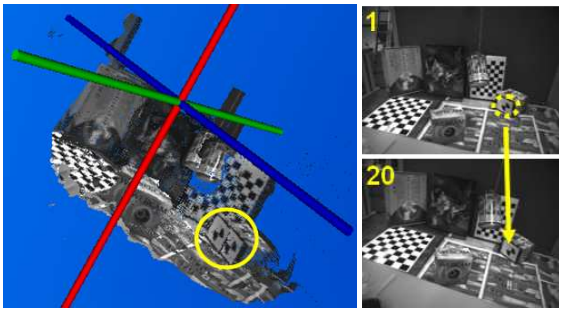

Figure 6. Overlaid rotated 3D depth maps from frames 1 and 20 (on the right) showing a clear mismatch, and circled image feature tracked to estimate translation.

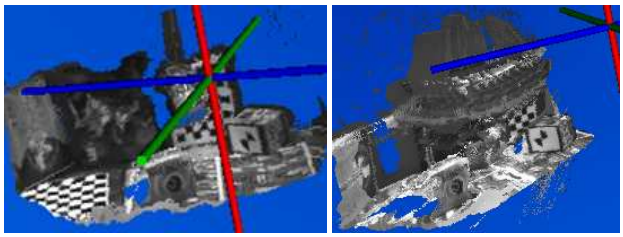

Figure 7. Depth maps rotated and translated to common world fixed frame of reference, for frames 1 and 20 on the left, and for full set of frames with moving pendulum on the right.

If the brightness changes smoothly with u, v and t, we can expand the left-hand-side by a Taylor series and reject the higher order terms to obtain

$$\nabla \boldsymbol{I} \cdot \frac{d\boldsymbol{u}}{dt} + \frac{\partial I}{\partial t} dt = 0 \tag{7}$$

where ∇I is the image gradient at pixel $\boldsymbol{u}$. These spatial and time derivatives can be estimated using a convolution kernel on the image frames.

But for each pixel we only have one constraint equation, and two unknowns. Only the *normal flow* can be determined, i.e., the flow along the direction of image gradient. The flow on the tangent direction of an isointensity contour cannot be estimated. This is the so called *aperture problem*. Therefore, to determine optical flow uniquely additional constraints are needed.

The problem is that a single pixel cannot be tracked, unless it has a distinctive brightness with respect to all of its neighbours. If a local window of pixels is used, a local constraint can be added, i.e., single pixels will not be tracked, but windows of pixels instead.

Barron *et al.* [4] present a quantitative evaluation of optical flow techniques, including the Lucas-Kanade method, that uses local consistency to overcome the aperture problem [12]. The assumption is made that a constant model can be used to describe the optical flow in a small window.

When the camera is moving and observing a static scene with some moving objects, some optical flow will be consistent with the camera ego-motion observing the static scene, other might be moving objects. Since the stereo provides a dense depth map, and we reconstruct camera motion, we can compute the expected projected optical flow in the image from the 3D data.

In the perspective camera model, the relationship between a 3D world point $\boldsymbol{x} = (X, Y, Z)^\mathsf{T}$ and its projection $\boldsymbol{u} = (u, v)^\mathsf{T}$ in the 2D image plane is given by

$$u = \frac{\boldsymbol{P}_1 (x, y, z, 1)^\mathsf{T}}{\boldsymbol{P}_3 (x, y, z, 1)^\mathsf{T}} \tag{8}$$

$$v = \frac{\boldsymbol{P}_2 (x, y, z, 1)^\mathsf{T}}{\boldsymbol{P}_3 (x, y, z, 1)^\mathsf{T}} \tag{9}$$

where $\boldsymbol{P}_j$ is the jth row of the camera projection matrix $\boldsymbol{P}$.

When the camera moves, the relative motion of the 3D point $\frac{d\boldsymbol{x}}{dt}$ will induce a projected optical flow given by

$$\frac{d\boldsymbol{u}_i}{dt} = \frac{\delta \boldsymbol{u}_i}{\delta \boldsymbol{x}} \frac{d\boldsymbol{x}}{dt} \tag{10}$$

where $\frac{\delta \boldsymbol{u}_i}{\delta \boldsymbol{x}}$ is the 2×3 Jacobian matrix that represents the differential relationship between $\boldsymbol{x}$ and $\boldsymbol{u}_i$, which can be obtained by differentiating (8) and (9).

Image areas where the computed flow is inconsistent with the expected one indicate moving objects, and the corresponding voxels can be segmented. This approach does not require the occupancy statistics memory, since it's differential and can be applied to pairs of successive frames.

Experimental results show that this method works on sequences with significant optical flow. However, this procedure is noise sensitive and, due to its differential based

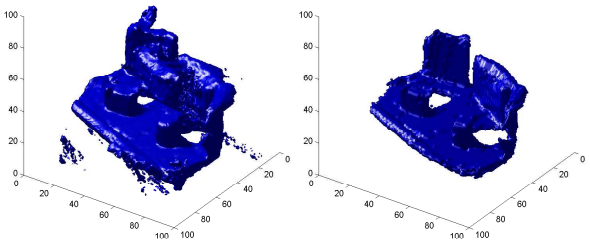

Figure 8. 3D volume of all accumulated voxels in frame sequence on the left, and with vote count above 30 on the right.

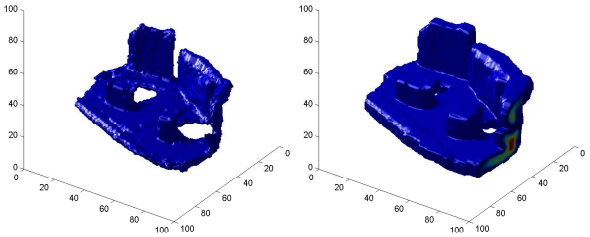

Figure 9. Background voxels after thinning for at least 6 neighbours on the left, and subsequent growth with a size 5^3 kernel on the right.

estimation, it performs poorly at low speeds, where the uncertainties in camera motion and optical flow are higher.

A summarising diagram of the procedures for both independent motion segmentation methods studied in this work is presented on Figure 4.

4. Results

The hardware system used to acquire data from a moving observer is shown in fig. 1. The stereo vision is provided by the Videre MEGA-D Digital Stereo Head [1], and the pose from the inertial and magnetic sensor package MT9-B from Xsens [2].

To compute range from stereo images we are using the SRI Stereo Engine with the Small Vision System (SVS) Software [6].

A scene was set up with a swinging cylindrical can to provide motion independent from the observer movement, as shown on Figure 5. The moving observer surveyed the scene performing map registration and subsequent independent motion segmentation as described below.

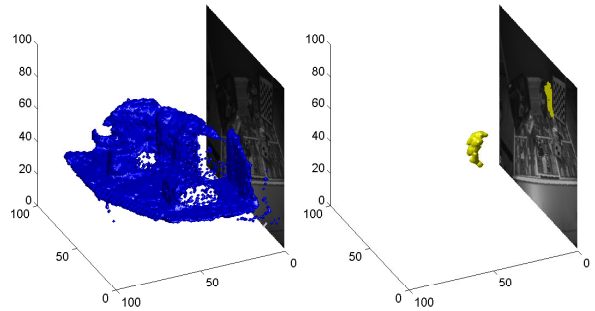

Figure 10. Initial voxel set and segmented moving object voxels for one frame.

4.1. Moving Depth Map Registration

As described above, the rotation update provided by the inertial and magnetic sensor package is applied to the successive depth maps. As shown on Figure 6, the depth maps are correctly rotated, but shifted due to the observer translation.

The translation was estimated by tracking an image feature, and observing the translation between the corresponding 3D points in the depth maps. Figure 6 shows data for frames 1 and 20 of a take of 200 frames with a moving observer of a static scene with a moving pendulum, for which the registration performed well.

The registered depth map can be seen in Figure 7. The fused map from frames 1 and 20 is shown on the left. On the right the fused map corresponding to the full set of frames is shown with the moving pendulum leaving its trace.

4.2. Background subtraction for Voxel Segmentation

The above results are shown with VRML rendering of the full set of computed points without voxel quantization. As described above, occupancy statistics can be used to identify the static scene voxels.

In a new test sequence, a one cubic meter volume of the observed space was chosen as the working volume, quantized to a $100 \times 100 \times 100$ array corresponding to $1\,cm^3$ voxels.

Figure 8 shows the 3D volume of all accumulated voxels for this test sequence with 130 frames, and the ones with a vote count above the empirically chosen threshold of 30. This choice was made based on the following observations: very low thresholds will mark slow objects as background; too high will segment newly observed static background as moving objects. Frame rate, observer motion and independent motion velocities are determining factors when choosing appropriate thresholds.

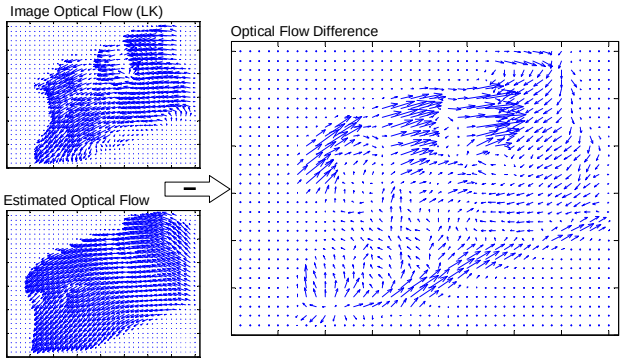

Figure 11. Difference between observed and estimated optical flow indicating areas inconsistent with static scene after camera motion compensation.

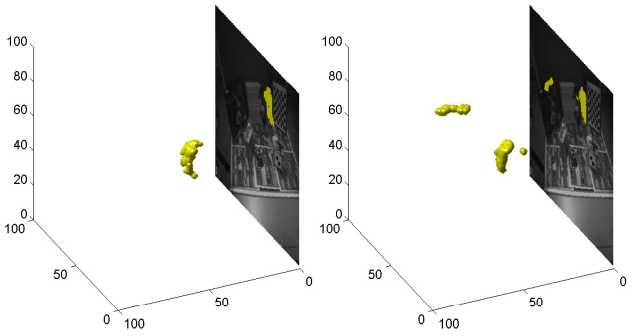

Figure 12. Output of the two methods for the same frame, voxel background subtraction on the left, and optical flow consistency on the right.

The result of thinning and growth filters applied to the background voxels is shown in Figure 9. Figure 10 shows initial voxel set and segmented moving object voxels after thinning and subsequent smoothing, for one frame from the above sequence. These results show that moving object voxels can be successfully segmented, however the moving observer has to cover the same scene more than once so that the background voxels can be correctly segmented.

4.3. Optical Flow Consistency Segmentation

Figure 11 shows the optical flow at one frame within a test sequence where the pendulum and observer were both moving. The image observed optical flow was estimated with the Lucas-Kanade [12] method applied to successive frames. The predicted flow was computed considering the 3D motion of the depth map relative to the camera, and projecting onto the image using (10).

The difference between the observed optical flow and the predicted flow indicate areas inconsistent with a static scene. The voxels associated with these image pixels correspond to moving objects. A decision threshold is applied to the optical flow difference to segment the voxels. The result for this frame is shown in figure 12 on the right.

The method works, but is clearly sensitive to noise due to the differential based estimation. In the complete test sequence there, frames with small camera motion performed poorly, since the uncertainties in camera motion and optical flow computation degrade the detection of moving objects.

5. Conclusions

Two methods were presented for motion segmentation for a moving observer of a background static scene with some independently moving objects. The moving observer has stereo vision and inertial and magnetic sensors to provide a rotation update. Having compensated rotation, translation can be obtained from a single tracked image feature. Depth maps from stereo vision can therefore be registered to a common frame of reference.

Occupancy statistics can then be used to segment the voxels between the static background scene and moving objects. However, the moving observer has to cover the same scene so that the background voxels can be correctly segmented. An alternative method is to check the consistency of the observed image optical flow. This approach is differential and can be applied to pairs of successive frames, but is more noise sensitive.

Figure 12 shows the output of the two methods for the same frame. The voxel background subtraction correctly identifies the independent motion. Due to this fact, in this work it also provided a ground truth to compare the optical flow consistency method. The optical flow consistency method also segments the independent motion, but with added false positives due to uncertainties in the optical flow computation and camera motion reconstruction.

On the other hand, voxel background subtraction requires a volumetric representation of the whole workspace, and also some past history statistics, which introduces a start-up lag of at least 10 frames, whereas optical flow consistency only needs the present and immediately preceding frames to function.

Therefore, a *hybrid* method can be devised which would take advantage of the strengths of *both* of these methods by using a differential approach based on optical flow whilst retaining a short-term memory of 3D space occupancy, since the inertial data allows fast depth map registration. Furthermore, this hybrid approach would more closely follow what indeed happens in biological/human perception systems,

where priors gathered from past states of the workspace being perceived are combined with fast low-level processing of retinal optical flow.

6. Discussion and Future Work

Although the techniques presented in this text are based on models that assume sensing technology that attempts to recreate the "hardware" of biological visuovestibular systems, no attempt has yet been made to follow the internal biological models of perception.

The usefulness of introducing models which mimic biological systems of perception and the limitations of biological perception posed by the physiological characteristics of biological motion sensors, which in certain situations yield partial or ambiguous information, has been demonstrated in previous research (see, for example, work by Reymond *et al.* [13]). Biological visuovestibular systems take into account ego-motion, and deal well with independent motion segmentation. In spite of this, however robust, biological perception estimation processes are prone to suffering from illusions, conflicts and ambiguities.

We have thus reached a point in which the next step will be to take artificial perception to the next level: *from bioinspired to biomimetic* — see figure 2.

We therefore propose in future work to perform psychophysical studies, such as in [13], of human visuovestibular models under a Bayesian framework, to implement these models as closely as possible using the technology presented on [9, 3, 8] in a robotic-based artificial perception system, to tackle 3D structure perception (specifically independent motion segmentation in the presence of self-motion), and to test the possibilities opened by the robustness of artificial sensor technology as opposed to biological sensory solutions on extreme perception tasks (see Figure 2). In the case of independent motion segmentation, we will address the use of inertial dynamic data to improve the optical flow consistency check, without depending on any tracked feature for the translation, and on combining the two methods to improve robustness.

Acknowledgements

This publication has been supported by EC-contract number *FP6-IST-027140, Action line: Cognitive Systems*. The contents of this text reflect only the author's views. The European Community is not liable for any use that may be made of the information contained herein.

References

[1] Videre Design. http://www.videredesign.com/.

[2] Xsens Technologies. http://www.xsens.com/.

[3] J. Alves, J. Lobo, and J. Dias. Camera-Inertial Sensor modeling and alignment for Visual Navigation. In *11th International Conference on Advanced Robotics*, pages 1693–1698, July 2003.

[4] J. Barron, D. Fleet, and S. Beauchemin. Performance of Optical Flow Techniques. *International Journal of Computer Vision*, 12(1):43–77, February 1994.

[5] C. Eveland, K. Konolige, and R. C. Bolles. Background modeling for segmentation of video-rate stereo sequences. In *Conference on Vision and Pattern Recognition*, Santa Barbara, CA, USA, June 1998.

[6] K. Konolige. Small vision systems: Hardware and implementation. In *Eighth International Symposium on Robotics Research*, Hayama, Japan, October 1997.

[7] R. Li and S. Sclaroff. Multi-scale 3d scene flow from binocular stereo sequences. In *To appear in Proc. IEEE Workshop on Motion and Video Computing*, January 2005.

[8] J. Lobo and J. Dias. Inertial Sensed Ego-motion for 3D Vision. In *International Conference on Advanced Robotics*, pages 1907–1914, July 2003.

[9] J. Lobo and J. Dias. Vision and Inertial Sensor Cooperation Using Gravity as a Vertical Reference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1597–1607, December 2003.

[10] J. Lobo and J. Dias. Inertial sensed ego-motion for 3d vision. *Journal of Robotic Systems*, 21(1):3–12, January 2004.

[11] J. Lobo and J. Dias. Relative pose calibration between visual and inertial sensors. In *ICRA 2005 Workshop on Integration of Vision and Inertial Sensors (InerVis2005)*, Barcelona, Spain, April 2005.

[12] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision . In *Proceedings of Imaging Understanding Workshop*, pages 674–679, 1981.

[13] G. Reymond, J. Droulez, and A. Kemeny. Visuovestibular perception of self motion modelled as a dynamic optimization process. *Biol. Cybern.*, 87:301–314, 2002.

[14] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 722–729, 1999.

[15] S. Vedula, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):475–480, March 2005.

[16] W. Yang, K. Ngan, J. Lim, and K. Sohn. Joint motion and disparity fields estimation for stereoscopic video sequences. *Signal Processing: Image Communication*, 20(3):265–276, March 2005.

[17] Y. Zhang and C. Kambhamettu. On 3-d scene flow and structure recovery from multiview image sequences. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 33(4):592–606, August 2003.